Supplementary material for Gavish, Nadler, Coifman: "Multiscale Wavelets on Trees, Graphs and High Dimensional Data", ICML 2010

A. Supplementary Material

In this section we present proofs of the various theorems in the paper. Recall that given a dataset X and its representation by a hierarchical tree, Eq. (5) defined a tree metric $d(x, y)$, whereas Eq. (6) defined (C, α)-Hölder smooth functions with respect to the tree metric. Let $f : X \to \mathbb{R}$. For any subset $Y \subset X$ we denote the mean and variance of f on Y as follows,

$$m(f, Y) = \frac{1}{|Y|} \sum_{x \in Y} f(x) \tag{16}$$

$$\sigma^2(f, Y) = \frac{1}{|Y|} \sum_{x \in Y} (f(x) - m(f, Y))^2. \tag{17}$$

Next, given the tree metric we denote by $B(x, r)$ the ball of radius r around x, that is

$$B(x, r) = \{y \in X \,|\, d(x, y) \le r\}$$

Observe that by definition, these balls are exactly the different folders of the tree that contain the node x.

The following lemma, standard in the theory of spaces of homogeneous type, will be useful in our proofs.

Lemma 1 *For any $x \in X$, $s > 0$ and $r > 0$ we have*

$$\int_{B(x,r)} d(x, y)^s d\nu(y) = \frac{1}{|X|} \sum_{y \in B(x,r)} d(x, y)^s \le C_s r^{s+1} \tag{18}$$

with $C_s = 2^{s+1} \left(1 - \frac{1}{2}\underline{B}\right) \le 2^{s+1}$.

Proof: Recall that by the definition of the tree metric, $d(x, y) \leqslant 1$, for any $x, y \in X$. Let $K \in \mathbb{N}$ be such that $2^{-K-1} < r \leqslant 2^{-K}$. Then

$$B(x, r) \subset \biguplus_{k=K}^{\infty} \left[B\left(x, 2^{-k}\right) \setminus B\left(x, 2^{-(k+1)}\right) \right]$$

Hence

$$\begin{aligned}
\int_{B(x,r)} d(x, y)^s d\nu(y) &\leqslant \sum_{k=K}^{\infty} \int_{B(x,2^{-k}) \setminus B\left(x,2^{-(k+1)}\right)} d(x, y)^s d\nu(y) \\
&\leqslant \sum_{k=K}^{\infty} \int_{B(x,2^{-k}) \setminus B\left(x,2^{-(k+1)}\right)} 2^{-ks} d\nu(y) \\
&\leqslant \sum_{k=K}^{\infty} 2^{-ks} \cdot \nu\left(B\left(x, 2^{-k}\right) \setminus B\left(x, 2^{-(k+1)}\right)\right) \\
&\leqslant \sum_{k=K}^{\infty} \left[2^{-ks} \left(2^{-k} - \underline{B} \cdot 2^{-(k+1)}\right)\right] ,
\end{aligned}$$

where the last inequality follows from the tree balance condition, Eq. (2). This gives

$$\begin{aligned}
\int_{B(x,r)} d(x, y)^s d\nu(y) &\leqslant \left(1 - \frac{1}{2} \cdot \underline{B}\right) \cdot \sum_{k=K}^{\infty} \left(\frac{1}{2^{s+1}}\right)^k \\
&\leqslant 2^{s+1} \left(1 - \frac{1}{2} \cdot \underline{B}\right) \left(2^{-K}\right)^{s+1} \leqslant 2^{s+1} \left(1 - \frac{1}{2} \cdot \underline{B}\right) r^{s+1} .
\end{aligned}$$

□.

Before proving theorem 1, we first introduce an alternative definition of function smoothness:

Definition A.1 *A function $f : X \to \mathbb{R}$ is (C, α)-Mean Hölder (w.r.t. the tree metric d) if for all $x \in X$ and any ball $B(x, r)$,*

$$\sigma\left(f, B(x,r)\right) \leqslant C \cdot \nu\left(B(x,r)\right)^{\alpha} \ . \tag{19}$$

where $\sigma(f, B(x,r))$ is defined in Eq. (17).

The following lemma shows that the two definitions of function smoothness w.r.t. the tree metric are related.

Lemma 2 *Let $f : X \to \mathbb{R}$ be (C, α)-Hölder with respect to the tree. Then f is $(2^{\alpha+1}C, \alpha)$ mean-Hölder.*

Proof of Lemma: Let $x \in X$ and let B be any ball around x. Since X is finite, for any $\varepsilon \geqslant 0$ small enough, we have $B = B(x, r)$ for $r = \nu(B) + \varepsilon$. Now,

$$\begin{aligned}
\int_B \left(f(x) - m(f,B)\right)^2 d\nu(x) &= \int_B \left(f(x) - \frac{1}{\nu(B)}\int_B f(y)\, d\nu(y)\right)^2 d\nu(x) \\
&= \frac{1}{\nu^2(B)}\int_B \left(\int_B f(x) - f(y)\, d\nu(y)\right)^2 d\nu(x) \leqslant \\
&\leqslant \frac{1}{\nu^2(B)}\int_B \left(\int_B |f(x) - f(y)|\, d\nu(y)\right)^2 d\nu(x) \, .
\end{aligned}$$

As $f : X \to \mathbb{R}$ is (C, α)-Hölder, this gives

$$\int_B \left(f(x) - m(f\,,\,B)\right)^2 d\nu(x) \leqslant \left(\frac{C}{\nu(B)}\right)^2 \int_B \left(\int_B d(x,y)^{\alpha}\, d\nu(y)\right)^2 d\nu(x) \, .$$

We now substitute $s = \alpha$ in Lemma 1 to obtain

$$\begin{aligned}
\int_B \left(f(x) - m(f\,,\,B)\right)^2 d\nu(x) &\leqslant \left(\frac{C}{\nu(B)}\right)^2 \int_B \left(2^{\alpha+1} r^{\alpha+1}\right)^2 d\nu(x) \\
&\leqslant \left(\frac{2^{\alpha+1}C}{\nu(B)}\right)^2 \nu(B)\, r^{2\alpha+2} \\
&\leqslant \left(\frac{2^{\alpha+1}C}{\nu(B)}\right)^2 \nu(B) \left(\nu(B) + \varepsilon\right)^{2\alpha+2} \ .
\end{aligned}$$

Since ε can be arbitrarily small, we conclude that

$$\int_B \left(f(x) - m(f\,,\,B)\right)^2 d\nu(x) \leqslant \left(\frac{2^{\alpha+1}C}{\nu(B)}\right)^2 \nu(B)^{2\alpha+3} = \left(2^{\alpha+1}C\right)^2 \nu(B)^{2\alpha+1}$$

and therefore

$$\sigma\left(f,\, B\right) = \sqrt{\frac{1}{\nu(B)}\int_B \left(f(x) - m(f\,,\,B)\right)^2 \nu(x)} \leqslant C 2^{\alpha+1} \nu(B)^{\alpha+1/2} \, . \tag{20}$$

Since $\nu(B) \leqslant 1$, the theorem follows. □.

Proof of Theorem 1: Recall that by definition, each Haar-like basis function $\psi_{\ell,k,j}$ is supported on the folder X_k^{ℓ}. It also has zero mean, namely $\int_{X_k^\ell} \psi_{\ell,k,j}(x)\, d\nu(x) = 0$, and unit norm, namely $\int_{X_k^\ell} \psi^2_{\ell,k,j}(x)\, d\nu(x) = 1$. Therefore,

$$\langle f\,,\,\psi_{\ell,k,j}\rangle = \int_{X_k^\ell} f(x)\psi_{\ell,k,j}(x)\, d\nu(x) = \int_{X_k^\ell} \left(f(x) - m(f, X_k^\ell)\right) \psi_{\ell,k,j}(x)\, d\nu(x) \, .$$

The Cauchy–Schwartz inequality now yields

$$\begin{aligned}
|\langle f\,,\,\psi_{\ell,k,j}\rangle| &\leqslant \sqrt{\int_{X_k^\ell} \left(f(x) - m\left(f\,,\,X_k^\ell\right)\right)^2 d\nu(x)} \cdot \sqrt{\int_{X_k^\ell} \left(\psi_{\ell,k,j}(x)\right)^2 d\nu(x)} \\
&= \sigma\left(f, X_k^\ell\right) \ .
\end{aligned}$$

According to Lemma 2, if f is (C,α) Hölder, it is $(C2^{\alpha+1},\alpha)$ mean-Hölder. In particular, Eq. (20) implies that

$$|\langle f\,\psi_{\ell,k,j}\rangle| \leqslant C2^{\alpha+1}\cdot\nu\left(X_k^\ell\right)^{\alpha+\frac{1}{2}}\ .$$

$\square$.

Proof of Theorem 2: Let $x,y\in X$ and let κ and λ be such that $folder(x,y)=X_\kappa^\lambda$. Our aim is to show that $|f(x)-f(y)|\leqslant C'\cdot\nu(X_\kappa^\lambda)^\alpha$ with C' given by Eq. (9).

To this end, we use the decomposition

$$f(x)=\sum_{\ell,k,j}\langle f,\psi_{\ell,k,j}\rangle\psi_{\ell,k,j}(x).$$

Note that by definition, for any coarse level $\ell<\lambda$ the samples x,y belong to the same folders, and thus $\psi_{\ell,k,j}(x)=\psi_{\ell,k,j}(y)$ for any k,j. Hence, the only terms contributing to the difference $f(x)-f(y)$ are those in the finer folders at levels $\ell=\lambda,\dots,L$, where x,y belong to *different* folders. That is,

$$\begin{aligned} f(x)-f(y) &= \sum_{\ell=\lambda}^{L}\sum_{j\in sub(\ell,\tau(\ell,x))}\left\langle f,\psi_{\ell,\tau(\ell,x),j}\right\rangle\cdot\psi_{\ell,\tau(\ell,x),j}(x)\\ &\quad-\sum_{\ell=\lambda}^{L}\sum_{j\in sub(\ell,\tau(\ell,y))}\left\langle f,\psi_{\ell,\tau(\ell,y),j}\right\rangle\cdot\psi_{\ell,\tau(\ell,y),j}(y)\end{aligned}$$

where $\tau(\ell,x)$ is the folder at level ℓ that contains x, $x\in X^\ell_{\tau(\ell,x)}$. Next, recall that by definition the functions $\psi_{\ell,k,j}$ are all normalized, and they are constant on all subfolders of X_k^ℓ. Thus,

$$\|\psi_{\ell,k,j}\|^2=\sum_{i\in sub(\ell,k)}\nu(X_i^{\ell+1})\psi^2_{\ell,k}(X_i^{\ell+1})=1$$

and so

$$|\psi_{\ell,k,j}(x)|\le\frac{1}{\sqrt{\nu(X_i^{\ell+1})}}\le\frac{1}{\sqrt{\underline{B}\,\nu(X_k^\ell)}}\,. \tag{21}$$

Combining the bound on $|\psi_{\ell,k,j}|$ with the bound on the coefficient decay of f gives that

$$\begin{aligned}|f(x)-f(y)| &\leqslant \frac{C}{\sqrt{\underline{B}}}\sum_{\ell=\lambda}^{L}\sum_{j\in sub(\ell,\tau(\ell,x))}\nu(X^\ell_{\tau(\ell,x)})^{\alpha+1/2}\frac{1}{\sqrt{\nu(X^\ell_{\tau(\ell,x)})}}\\ &+\frac{C}{\sqrt{\underline{B}}}\sum_{\ell=\lambda}^{L}\sum_{j\in sub(\ell,\tau(\ell,y))}\nu(X^\ell_{\tau(\ell,y)})^{\alpha+1/2}\frac{1}{\sqrt{\nu(X^\ell_{\tau(\ell,y)})}}\end{aligned} \tag{22}$$

Finally, since the tree is balanced, $\nu(X^\ell_{\tau(\ell,x)})\le\overline{B}^{\ell-\lambda}\nu(X^\lambda_\kappa)$, and $|sub(\ell,k)|\le\frac{1}{\underline{B}}-1$. Thus,

$$\begin{aligned}|f(x)-f(y)| &\leqslant \frac{2C(1-\underline{B})}{\underline{B}^{3/2}}\sum_{\ell=\lambda}^{L}\left(\overline{B}^\alpha\right)^{\ell-\lambda}\nu(X^\lambda_\kappa)^\alpha\\ &\leqslant \frac{2C}{\underline{B}^{3/2}}\frac{1}{1-\overline{B}^\alpha}\nu(X_\kappa^\lambda)^\alpha=C'\nu(X_\kappa^\lambda)^\alpha\,.\end{aligned}$$

$\square$.

Proof of Theorem 3: Let $\hat f=\sum_{|I|>\epsilon}a_I h_I(x)$. Then

$$\begin{aligned}\|f-\hat f\|_1 &= \sum_x|f(x)-\hat f(x)|=\sum_x\left|\sum_{|I|<\epsilon}a_I h_I(x)\right|\\ &\leqslant \sum_{|I|<\epsilon}|a_I|\sum_{x\in I}|h_I(x)|\end{aligned} \tag{23}$$

but according to the assumptions of the theorem, $|h_I(x)| \leqslant 1/|I|^{1/2}$ and $supp(h_I) = |I|$. Hence, $\sum_{x\in I} |h_I(x)| < \epsilon/\sqrt{\epsilon} = \sqrt{\epsilon}$. Combining this with the entropy condition on the coefficients, $\sum_I |a_I| \leqslant C$ the theorem follows. □

Proof of Theorem 4: Recall that the coefficient $\hat{a}_{\ell,k,j}$ is given by Eq. (12) if all subfolders of X_k^ℓ at level $\ell+1$ each contain at least one labeled point. Otherwise, $\hat{a}_{\ell,k,j}$ is set to zero. Denote by R the event that at least one of the subfolders of X_k^ℓ does not contain labeled points. First of all,

$$\begin{aligned}\Pr[R] &\leqslant \sum_{i\in sub(\ell,k)} \Pr\left[\,|S\cap X_i^{\ell+1}|=0\right] = \sum_{i\in sub(\ell,k)} (1-\nu(X_i^{\ell+1}))^{|S|} \\ &\leqslant \sum_{i\in sub(\ell,k)} e^{-|S|\nu(X_i^{\ell+1})} \leqslant \frac{1}{\underline{B}} e^{-|S|\underline{B}\nu(X_k^\ell)}\end{aligned}$$

Conditional on the event R, we have $\mathbb{E}[\hat{a}_{\ell,k,j}] = var[\hat{a}_{\ell,k,j}] = 0$, whereas under R^c, we have that $\mathbb{E}[\hat{a}_{\ell,k,j}] = a_{\ell,k,j}$, and after some algebraic manipulations,

$$var[\hat{a}_{\ell,k,j}\mid R^c] = \sum_{i\in sub(\ell,k)} \nu^2(X_i^{\ell+1})\psi^2_{\ell,k,j}(X_i^{\ell+1})\frac{\sigma^2(f,X_i^{\ell+1})}{|S\cap X_i^{\ell+1}|} \tag{24}$$

To compute the mean squared error of the estimator $\hat{a}_{\ell,k,j}$ we use the identity

$$\mathbb{E}\left[\hat{a}_{\ell,k,j}-a_{\ell,k,j}\right]^2 = var[\hat{a}_{\ell,k,j}] + \left(\mathbb{E}[\hat{a}_{\ell,k,j}]-a_{\ell,k,j}\right)^2 . \tag{25}$$

Regarding the second term in (25), we have that $\mathbb{E}[\hat{a}_{\ell,k,j}] = a_{\ell,k,j}\,(1-\Pr[R])$. Thus,

$$\left(\mathbb{E}[\hat{a}_{\ell,k,j}]-a_{\ell,k,j}\right)^2 \;=\; a^2_{\ell,k,j}\Pr[R]^2 . \tag{26}$$

As for the first term in (25), let Z be the random variable defined as the indicator function of the event R, $Z = \mathbf{1}_R$. By the variance decomposition formula

$$var\left[\hat{a}_{\ell,k,j}\right] = \mathbb{E}\left[var\left[\hat{a}_{\ell,k,j}\middle|Z\right]\right] + var\left[\mathbb{E}\left[\hat{a}_{\ell,k,j}\middle|Z\right]\right] \tag{27}$$

Now, by (24),

$$\mathbb{E}\left[var[\hat{a}_{\ell,k,j}\middle|Z]\right] = \Pr[R^c]\sum_{i\in sub(\ell,k)} \nu^2(X_i^{\ell+1})\psi^2_{\ell,k,j}(X_i^{\ell+1})\sigma^2(f,X_i^{\ell+1})\mathbb{E}\left[\frac{1}{|S\cap X_i^{\ell+1}|}\middle| R^c\right]$$

For $|S|\gg 1$, we approximate the conditioning on R^c by the (simpler) conditioning on $\{|S\cap X_i^{\ell+1}|>0\}$. This gives

$$\mathbb{E}\left[var[\hat{a}_{\ell,k,j}\middle|Z]\right] \;=\; \Pr[R^c]\sum_{i\in sub(\ell,k)} \nu^2(X_i^{\ell+1})\psi^2_{\ell,k,j}(X_i^{\ell+1})\sigma^2\left(f,X_i^{\ell+1}\right)\frac{\mathbb{E}\left[A_i\right]}{\Pr[\,|S\cap X_i^{\ell+1}|>0\,]} \tag{28}$$

where

$$A_i = \begin{cases} \frac{1}{|S\cap X_i^{\ell+1}|} & |S\cap X_i^{\ell+1}|>0 \\ 0 & |S\cap X_i^{\ell+1}|=0 \end{cases} .$$

The quantity $\mathbb{E}\left[A_i\right]$ is known as the first *inverse moment* of the Binomial distribution $Bin\left(|S|,\, \nu(X_i^{\ell+1})\right)$. Asymptotic expansions of this quantity have been studied extensively. In Rempala (2003), it was proved that

$$\mathbb{E}\left[A_i\right] = \frac{1}{|S|\cdot\nu(X_i^{\ell+1})} + o\left(\frac{1}{|S|}\right) .$$

Using this approximation in (28) gives, up to an $o(1/|S|)$ error

$$\mathbb{E}\left[var\left[\hat{a}_{\ell,k,j}\middle|Z\right]\right] \;\approx\; \frac{\Pr[R^c]}{|S|}\sum_{i\in sub(\ell,k)} \nu(X_i^{\ell+1})\psi^2_{\ell,k,j}(X_i^{\ell+1})\frac{\sigma^2\left(f,X_i^{\ell+1}\right)}{\Pr[\,|S\cap X_i^{\ell+1}|>0]} .$$

As f is (C,α)-Hölder, according to Lemma 2 it is (C_1,α) mean-Hölder with $C_1 = 2^{\alpha+1}C$. Thus, $\sigma^2(f, X_i^{\ell+1}) \leqslant C_1^2 \nu(X_i^{\ell+1})^{2\alpha}$. Since the tree is balanced, $\nu(X_i^{\ell+1}) \leq \overline{B}\nu(X_k^\ell)$. In addition,

$$\frac{1}{\Pr[\,|S \cap X_i^{\ell+1}| > 0]} \leqslant \frac{1}{1 - e^{-|S|\nu(X_i^{\ell+1})}} \leqslant \frac{1}{1 - e^{-|S|\underline{B}\nu(X_k^\ell)}}$$

Therefore,

$$\begin{aligned}
\mathbb{E}\left[var\left[\hat{a}_{\ell,k,j}\middle|Z\right]\right] &\leqslant \frac{1}{|S|}\frac{C_1^2\overline{B}^{2\alpha}\nu^{2\alpha}(X_k^\ell)}{1 - e^{-|S|\underline{B}\nu(X_k^\ell)}}\sum_{i\in sub(\ell,k)} \nu(X_i^{\ell+1})\psi_{\ell,k,j}^2(X_i^{\ell+1}) \\
&= \frac{1}{|S|}\frac{C_1^2\overline{B}^{2\alpha}\nu^{2\alpha}(X_k^\ell)}{1 - e^{-|S|\underline{B}\nu(X_k^\ell)}}\,. \qquad (29)
\end{aligned}$$

where the summation is simply $\|\psi_{\ell,k,j}\|^2 = 1$.

For the second term in Eq. (27), note that

$$\mathbb{E}\left[\hat{a}_{\ell,k,j} \mid Z\right] = \begin{cases} a_{\ell,k,j} & \text{under } R^c \\ 0 & \text{under } R \end{cases} \qquad (30)$$

Therefore,

$$var\left[\mathbb{E}\left[\hat{a}_{\ell,k,j}\middle|Z\right]\right] = a_{\ell,k,j}^2\left(1 - \Pr[R]\right)\Pr[R]. \qquad (31)$$

Combining (29), (31) into (25) gives that

$$\begin{aligned}
\mathbb{E}[\hat{a}_{\ell,k,j} - a_{\ell,k,j}]^2 &\leqslant \frac{1}{|S|}\frac{C_1^2\overline{B}^{2\alpha}\nu^{2\alpha}(X_k^\ell)}{1 - e^{-|S|\underline{B}\nu(X_k^\ell)}} + a_{\ell,k,j}^2\left(1 - \Pr[R]\right)\Pr[R] + a_{\ell,k,j}^2\Pr[R]^2 \\
&\leqslant \frac{1}{|S|}\frac{C_1^2\overline{B}^{2\alpha}\nu^{2\alpha}(X_k^\ell)}{1 - e^{-|S|\underline{B}\nu(X_k^\ell)}} + \frac{1}{\underline{B}}e^{-|S|\underline{B}\nu(X_k^\ell)} \cdot a_{\ell,k,j}^2\,. \qquad (32)
\end{aligned}$$

Finally, to prove the formula for the mean squared error in estimating f we note that due to the orthogonality of the Haar-like basis functions,

$$\begin{aligned}
\mathbb{E}\left\|f - \hat{f}\right\|^2 &= \mathbb{E}\left[\left\|\sum_{\ell,k,j}\left(a_{\ell,k,j} - \hat{a}_{\ell,k,j}\right)\psi_{\ell,k,j}\right\|^2\right] = \mathbb{E}\left[\sum_{\ell,k,j}\left(a_{\ell,k,j} - \hat{a}_{\ell,k,j}\right)^2\right] \\
&= \sum_{\ell,k,j}\mathbb{E}\left[a_{\ell,k,j} - \hat{a}_{\ell,k,j}\right]^2.
\end{aligned}$$

Hence,

$$\begin{aligned}
\mathbb{E}\,\|f - \hat{f}\|^2 &\leqslant \frac{C_1^2\overline{B}^{2\alpha}}{|S|}\sum_{\ell,k,j}\frac{\nu(X_k^\ell)^{2\alpha}}{1 - e^{-|S|\underline{B}\nu(X_k^\ell)}} + \frac{1}{\underline{B}}\sum_{\ell,k,j}e^{-|S|\underline{B}\nu(X_k^\ell)}a_{\ell,k,j}^2 \\
&\leqslant \frac{C_1^2\overline{B}^{2\alpha}}{|S|}\sum_{\ell,k,j}\frac{\left(\overline{B}^{2\alpha}\right)^{\ell-1}}{1 - e^{-|S|\underline{B}^\ell}} + \frac{2^{2\alpha+1}C_1^2}{\underline{B}}\sum_{\ell,k,j}e^{-|S|\underline{B}^\ell}\left(\overline{B}^{2\alpha+1}\right)^{\ell-1} \qquad (33)
\end{aligned}$$

□